Learning from demonstration of trajectory preferences through causal modeling and inference

Learning from demonstration of trajectory preferences through causal modeling and inference

Daniel Angelov
School of Informatics
The University of Edinburgh
d.angelov@ed.ac.uk

Subramanian Ramamoorthy
School of Informatics
The University of Edinburgh
s.ramamoorthy@ed.ac.uk

***Abstract*—Learning from demonstration is associated with acquiring a solution to a task by mimicking a teacher demonstrator. Understanding the underlying reasons and in turn preferences that lead to a demonstration can yield better task comprehension. We present a generative model that describes a table-top task in terms of a causal model with respect to known concepts (e.g., the notion of a fork). Causal reasoning in the latent space of this generative model fully describes the meaning of the demonstration, e.g., that we would like to move far away from the fork. We show that by sampling from the model latent space, we can learn a solution to the problem that defines the task being demonstrated. We use a simulated kitchen table-top environment to show changes in the underlying trajectory preference of demonstrations for different objects. The ability to generate additional data through introspection of the latent space allows us to confirm the causal model for the problem.**

I. Introduction

As we move from robots dedicated to specific pre-programmed tasks to more general purpose tasks, there is a need for easy re-programmability of these robots. A promising approach to such easy re-programming is learning from demonstration, i.e., by enabling the robot to mimic behaviours shown to it by a human expert.

With such a setup, we can abstract away from having to handcraft rules, and allow the robot to learn by itself, including the preferences exhibited by the teacher within the demonstration. Often these innate preferences are not explicitly articulated and are mostly biases resulting from experiences with other unrelated tasks sharing parallel environmental corpora - Figure. 1.1. The ability to notice, understand and reason causally about these deviations, whilst learning to perform the shown task is of high interest.

Similarly, other methods for learning from demonstration (LfD) as discussed by Argall et al. [1] are focused in finding a general mapping from observed state to an action, thus modeling the system or attempting to capture the hight level user intentions into a plan. The resulting policies are not generally used as *generative models*. And as highlighted by Sünderhauf et al. [15] one of the fundamental challenges with robotics is the ability to reason about the environment, beyond a state-action mapping.

Thus, when receiving a positive demonstration, we should aim to understand the causal reasons differentiating it from a non-preferential one, rather than purely learning the particular trajectory. When people demonstrate a concept, they rarely refer to the specific trajectory alone, but rather a set of trajectories that display particular features. In other words, we want to find groups of trajectories with similar characteristics that may be viewed as clusters. We are interested in learning these clusters and their boundary, so that subsequent new trajectories can be classified according to whether they are good representatives of the class of feasible behaviours.

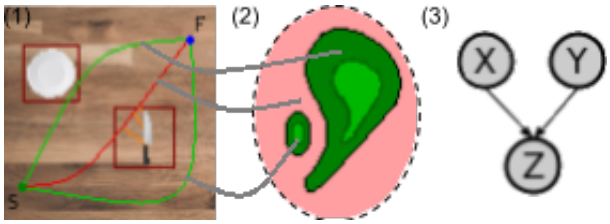

Fig. 1: **(1)** Demonstrations that satisfy the human thought behind their preference of using trajectories that pass further away from dangerous objects (i.e. a knife), but do not have problems moving close to others (i.e. plates). **(2)** A specific environment can have multiple clusters of varying distributions of valid trajectories in its latent space. **(3)** Underlyingly, the validity of trajectories with respect to the human preference can be represented as a causal model. Whether a trajectory is part of a cluster (Z) is affected by the specific path (X) and the environment (Y).

It is often the case that in problems that allow for great flexibility in their solution, different experts may generate solutions that are part of different clusters - Figure. 1.2. In cases where we naively attempt to perform statistical analysis, we may end up collapsing to a single mode, or merging the modes in a semantically senseless manner (i.e. averaging trajectories for going left/right around an object).

We present a method for introspecting in the latent space which allows us to relax some of the assumptions illustrated above and more concretely to:

- find valid, varied solutions of a problem through sampling a generative model, which we learn
- show boundaries of the valid clusters of the demonstration in its latent space
- use those boundaries and given key environmental features within the demonstration to counterfactually reason

about the underlying feature preference of the demonstration and build a causal model describing it.

The method relies on generating a latent space from unlabeled environmental observations with the demonstrator's trajectories. The teacher's positive and negative examples are used as a guide for estimating the preferences and validity of the trajectory parametrization.

In the following section we will show relevant work in the field of learning from demonstration as well as causality. In the next section we will discuss the methodology for building a model to describe the mental preference of the demonstrator. It is followed by the capability of extracting a structured causal model by counterfactual reasoning. We show our results and add concluding notes.

II. Related Work

A. Learning from Demonstration

Learning from demonstration has involved a variety of different methods for approximating the policy. In some related work, the state space is partitioned and the problem is viewed as a classification approach. This allows for the environment state to be in direct control of the robot and to command its discrete actions - using Neural Networks (J Matari'c [10]), Bayesian Networks (Inamura [9]), Gaussian Mixture Models (Chernova and Veloso [3]). Alternatively, it can be used to classify the current step in a high level plan Thomaz and Breazeal [16] and execute predetermined low level control.

In cases where a continuous action space is preferred, regressing from the observation space can be achieved by methods like Locally Weighted Regression Cleveland and Loader [4].

Robotisists e.g., Sünderhauf et al. [15], have long viewed that reasoning as part of planning is dependent on reasoning about objects, semantics and their geometric manifestations. This process is based on the view that structure within the demonstration should be exploited to better ground symbols between modalities and to the plan.

B. Causality and state representation

The variability of environmental factors makes it hard to build systems relying only on correlation data statistics for specifying their state space. Methods that rely on causality, Pearl [12], Harradon et al. [7], and learning the cause and effect structure, Rojas-Carulla et al. [14], are much better suited to support the reasoning capabilities for transferring core knowledge between situations. Interacting with the environment allows robots to perform manipulations that can convey new information to update the observational distribution or change their surrounding and in effect perform interventions within the world.

Learning sufficient state features has been highlighted by Argall et al. [1] as a future challenge for the LfD community. The problem of learning disentangled representations aims to generate a good composition of a latent space, separating the different modes of variation within the data. Higgins et al. [8], Chen et al. [2] have showed promising improvements in disentangling of the latent space with minimal or no assumption by manipulating the Kullback - Leibler divergence loss of a variational auto encoder. Denton and Birodkar [5] shows how the modes of variation for content and temporal structure should be separated and can be extracted to improve the quality of the next frame video prediction task, if temporal information is added as a learning constraint. While the disentangled representations may not directly correspond to the factors defining action choices, Johnson et al. [11] adds a factor graph and composes latent graphical models with neural network observation likelihoods.

The ability to manipulate the latent space and separate variability as well as obtain explanation about behavior is also of interest to the interpretable machine learning field, as highlighted by Doshi-Velez and Kim [6].

III. Experimental Setup

In this section we illustrate how modeling the preferences of a human demonstrator's trajectories, in a table-top manipulation scenario within a neural network model, can be later used to infer causal links through a set of known features about the environment.

A. Dataset

The environment chosen for the experiment consists of a top down view of a tabletop on which a collection of items, O=$\{$*knife, plate*$\}$, usually found in a kitchen environment, have been randomly distributed. The task that the demonstrator has to accomplish is to move a robotic arm from one end on the table to the opposite (bottom left to top right) by demonstrating a trajectory, whilst encompassing any human preferences around the set of objects they may have.

The teacher is given a *100x100* input image I as shown on Figure. 2 and has to produce a number of possible trajectories, some that satisfy the solution and some that break the demonstrators preferences - Figure. 1.1. We describe the trajectories using a Bezier curve representation and the parameterization is the location of the central control point parametrized by θ, with the first at the start of the trajectory, and the last - at the end. In the current setup, the demonstrator is implemented as an expert system that has the human preference encoded into its rules.

The dataset consists of 1000 randomly generated scenes consisting of between 1 and 2 objects and 10 example trajectory per scene.

Fig. 2: Examples of possible scene configurations.

B. Preference Model

The main capability that we want from our model is to structure the latent space in a way that would not only allow us to solve and generate good trajectories, but also to manage the variability that would be needed to estimate the causal link between valid trajectories and the world representation.

We use a convolutional variational auto-encoder to compress the world representation *I* to a latent space z_I, disjoint from the parameterization of the trajectories z_θ. The full latent space is modeled as the concatenation of the world space and trajectory space $z = z_I \cup z_\theta$ as seen on Figure. 3.

In order to better shape the latent space, we add a β term to the KL divergence loss as in Higgins et al. [8]. Additionally, we add an extra preference binary cross-entropy loss (scaled by γ) associated with the ability of the full latent space z in predicting whether that trajectory with the associated world satisfy the preference of the demonstrator - v. The full loss can be seen in Eq. 1.

$$\begin{aligned}\mathcal{L}(\theta,\phi;I,&z_I,z_\theta,v,\alpha,\beta,\gamma) = \\ &\alpha\mathbb{E}_q(z_I|I)[log_p(I|z_I)] \\ &-\beta D_{KL}(q_\phi(z_I|I)||p(z_I)) \\ &+\gamma\left[v\log(C(z))+(1-v)\log(1-C(z))\right]\end{aligned} \tag{1}$$

We evaluate the performance of the model on its ability to correctly predict the preference exhibited by the demonstrator.

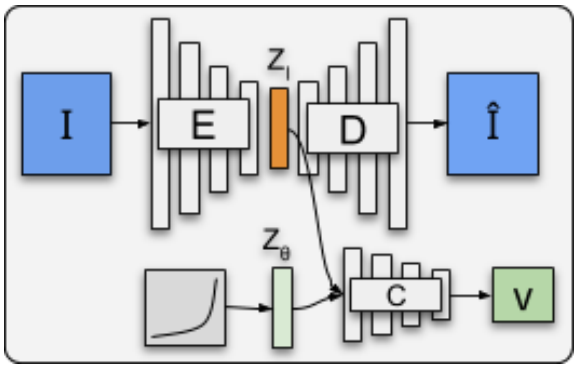

Fig. 3: Preference model architecture. The environmental image *I* is passed through an Encoder-Decoder Convolutional Network, with a 16-8-4 number of 3x3 convolutions, followed by fully connected layer, to create a compressed representation $Z_I, Z_I \in R^{15}$. It is passed along with the trajectory parameterization $Z_\theta, Z_\theta \in R^2$ through a 3-layer fully connected classifier network that checks the validity of the trajectory (*C(z)*) with respect to the mental model behind the human preference.

IV. Causal Modeling

Naturally, we can examine our causal understanding of the environment only with the limited set of features, O, that we can comprehend about the world. We work under the assumption that an object detector is available for these objects (as the focus of this work is on elucidating the effect of these objects on the trajectories rather than on the lower level computer vision task of object detection per se). Given this, we can construct specific world configurations to test a causal model and use the above learned preference model as a surrogate to inspect the validity of proposed trajectories.

If we perform a search in the latent space z_θ, we can find boundaries of trajectory validity as shown on Figure. 6. We can intervene and counterfactually alter parameters of the environment and see the changes of the trajectory boundaries. By looking at the difference of boundaries in simple cases where we can test for associational reasoning, we can infer whether humans have different preferences between the items.

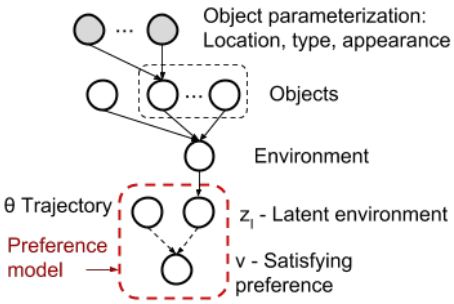

Fig. 4: We assume that the environment, compressed to z_I, is composed of objects, some of which parameters are known. A trajectory is parameterized by θ, which alongside the factors z_I and v are part of the preference model.

We are interested in establishing the causal relationship within the preference model as shown on Figure. 4. We define our Structural Causal Model (SCM), following notation of Peters et al. [13] as

$$\mathfrak{C} := (\mathbf{S}, P_\mathbf{N}),\ S = \{X_j := f_j(\mathbf{PA}_j, N_j)\}$$

where nodes $\mathbf{X} = \{Z_\theta, Z_I, V\}$ and $\mathbf{PA}_j = \{X_1, X_2, ..X_n\}\backslash\{X_j\}$. Given some observation $\mathbf{x}$, we can define a counterfactual SCM $\mathfrak{C}_{\mathbf{X}=\mathbf{x}} := (\mathbf{S}, P_\mathbf{N}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}})$, where $P_\mathbf{N}^{\mathfrak{C}|\mathbf{X}=\mathbf{x}} := P_{N|\mathbf{X}=\mathbf{x}}$

We can choose a specific trajectory $z_\theta \sim p(Z_\theta)$ and environment $z_I \sim p(Z_I)$ and use the preference model and confirm the causal links $Z_I \to v$ and $Z_\theta \to v$ by showing:

$$\mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}\right] \neq \mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(Z_\theta:=z_\theta)}\right] \tag{2}$$

$$\mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}\right] \neq \mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(Z_I:=z_I)}\right] \tag{3}$$

V. Results

The preference model was trained with the full loss as described in Eq. 1, with $\alpha = 1, \beta = 10, \gamma = 1e7$. We compare it to a model with the same hyper parameters with a single example trajectory per scene, and a truncated loss, leaving only the classification cross-entropy term $\alpha = 0, \beta = 0, \gamma = 1$ similar to most LfD classification methods. The accuracy on a test set is shown on Figure. 5. The single trajectory cases show that adding a β-VAE loss improves the performance to 64.8% (vs 60.2%). The full model reached 92.4%.

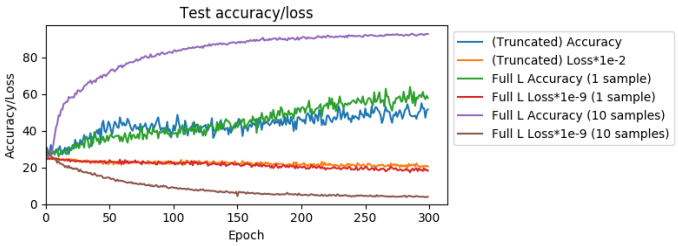

Fig. 5: Accuracy/Classification Loss behavior during training on a test dataset. The learning rate (Adam, $\alpha = 10^{-3}$) was annealed in all cases by 0.95 every 20 epochs.

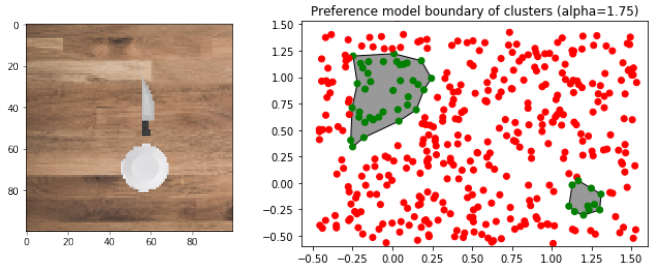

Fig. 6: On the right showing 400 samples for $z_\theta \sim p(Z_\theta)$ trajectories which are colored by compliance to the meaning a teacher has used when demonstrating (green for valid), given a specific latent z_I, which is illustrated on the left. The gray boundary is generated by fitting an alpha shape to the points. The two clusters are describing trajectories that can be summarized as "trying to stay away from the knife going above it" and "try to go below the plate".

The boundaries of the teacher preference within a specified environment can be seen on Figure. 6. We can observe samples from different trajectory parameterizations and their validity. In the current instance they are clustered following a preference to traverse in the top left part of the physical space.

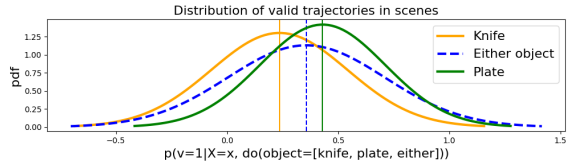

Fig. 7: Shows different distributions associated with valid trajectories with different singular objects in a world. The expected value of valid trajectories differs between changing the objects to [0.2361, 0.3571, 0.4303] corresponding to [knife, either objects, plate]. It indicates that the probability of sampling a valid trajectory with the world containing a plate is higher than when a knife is present, indicating ease of finding a preferred trajectory.

In the simple experiment, where a single item is placed around the environment, we can see the expectations and variance of the validity of a trajectory on Figure. 7. The distributions are different indicating the human demonstrations has exhibited different preferences in choosing a rational trajectory with the items being part of the environment.

In the counterfactual case, we evaluate Eq.2 and 3 over a set of $1e5$ random latent space points and obtain that:

$$\mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x}}\right] = 0.26 \tag{4}$$

$$\mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(Z_\theta:=z_\theta)}\right] = 0.20 \tag{5}$$

$$\mathbb{E}\left[P_v^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(Z_I:=z_I)}\right] = 0.34 \tag{6}$$

This indicates that indeed Z_I and Z_θ causally change the validity of a trajectory. The final SCM is shown on Figure. 8.

Relaying on the SCM we can further answer questions as: What are possible valid trajectories, given an environment configuration I? What world configurations may make a trajectory θ as a valid?

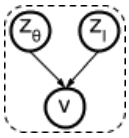

Fig. 8: Final structural causal model. The validity of the preference of a trajectory is the effect of both the environment and the specified trajectory.

VI. Conclusion

We show how in a learning from demonstration setup, we can create a generative model for the preference teachers exhibit whilst performing a task. Through sampling, the model can be used to both solve a particular environment to show a trajectory respecting the human intuition, as well as a proxy to estimate the causal relationship between the environment, trajectory and its validity.

In future work we want to analyse what aspects of the objects cause the validity of the trajectory and by more extensive parametrization of the path, find which parts of them are breaking the expectations. Additionally, having a causal approach to requesting demonstrations, we want to lower the number of needed examples.

Acknowledgments

This research is supported by the Engineering and Physical Sciences Research Council (EPSRC), as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh. Grant reference EP/L016834/1.

References

[1] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57 (5):469 – 483, 2009.

[2] T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. *ArXiv e-prints*, February 2018.

[3] Sonia Chernova and Manuela Veloso. Confidence-based policy learning from demonstration using gaussian mixture models. In *Proceedings of the 6th International*

Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, 2007.

[4] William S. Cleveland and Clive R. Loader. Smoothing by local regression: Principles and methods.

[5] E. Denton and V. Birodkar. Unsupervised Learning of Disentangled Representations from Video. *ArXiv e-prints*, May 2017.

[6] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints*, February 2017.

[7] M. Harradon, J. Druce, and B. Ruttenberg. Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations. *ArXiv e-prints*, February 2018.

[8] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework.

[9] T. Inamura. Acquisition of probabilistic behavior decision model based on the interactive teaching method. *Proc. 9th Int'l Conf. on Advanced Robotics*, pages 523–528, 1999. URL https://ci.nii.ac.jp/naid/20000105704.

[10] Maja J Matari'c. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. *MIT Press, Cambridge, MA, USA*, 11 1999.

[11] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams. Composing graphical models with neural networks for structured representations and fast inference. *ArXiv e-prints*, March 2016.

[12] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 052189560X, 9780521895606.

[13] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[14] M. Rojas-Carulla, M. Baroni, and D. Lopez-Paz. Causal Discovery Using Proxy Variables. *ArXiv*.

[15] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke. The Limits and Potentials of Deep Learning for Robotics. *ArXiv e-prints*, April 2018.

[16] A. Thomaz and C. Breazeal. Tutelage and socially guided robot learning. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE)*.